Network type: Peer to peer network

Characteristics:

* "peers" are computer systems which are connected to each other via the Internet.
* Files can be shared directly between systems on the network without the need of a central server.
* In other words, each computer on a P2P network becomes a file server as well as a client.

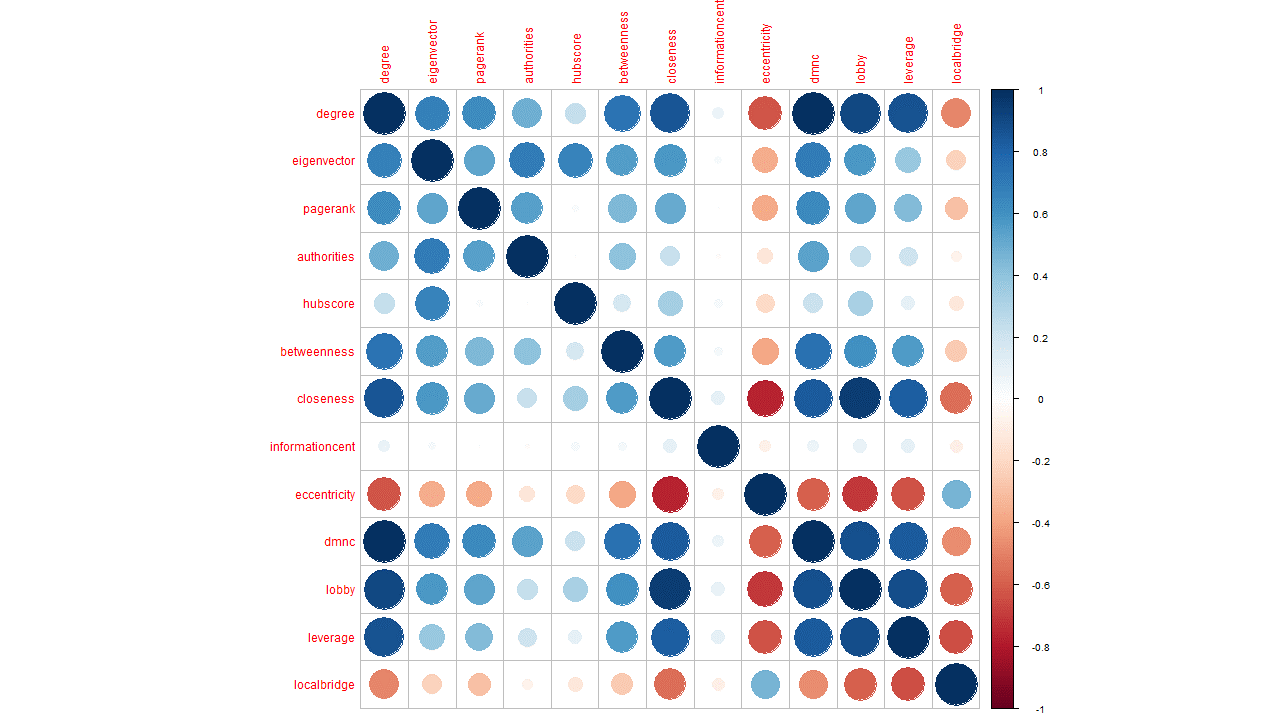
Assumptions about P2P: What defines important nodes? What separates influential nodes from non-influential ones- where does the variance occur?

* Seeders are more important- almost all nodes are seeders except a few. The ones that seed a bigger variety of demanded files, or a bigger number of demanded files, can be marked as influential.
* Nodes who seed many files will be connected to many different nodes. Nodes that share a variety of file types will also be connected to many nodes, hence they are expected to have higher degree.
* Influential nodes are likely to have relatively higher degree than their neighbors ( for example someone who seeds anime, movies, music and games is supposed to have higher degree than its neighbors who seed only anime or movies). Hence influential nodes are likely to have higher leverage values as well.
* Nodes that seed many files will also be closer to other nodes as they are more likely to be directly connected to more nodes, therefore they are expected to have higher closeness values.
* Nodes that are connected to other nodes of higher degree, are likely to also seed files that are in demand. Hence even if they have low degree themselves, they will be contributing to overall information transfer. This makes eigenvector centrality an important factor.

Results:

Peer-to-peer (P2P) networks are networks in which all nodes have the possibility of being connected to each other. This is due to the nature of the network, as edges are formed between nodes based on the availability of files and file segments. Users of P2P networks often share all kinds of files in these networks, hence it is one big community where all nodes can potentially connect to each other, rather than small cliques or closely connected communities.

Correlation plot:



PCA:

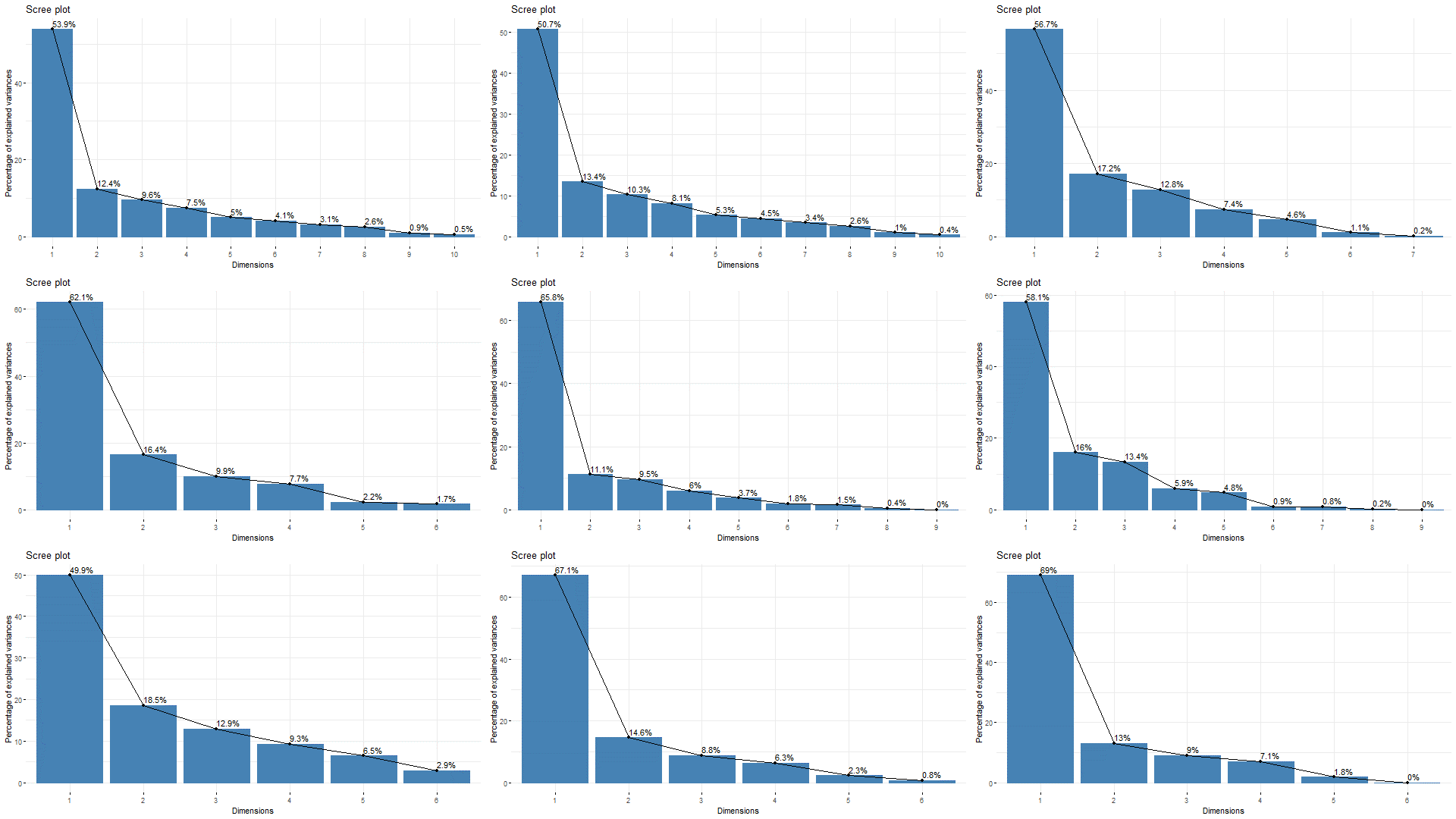
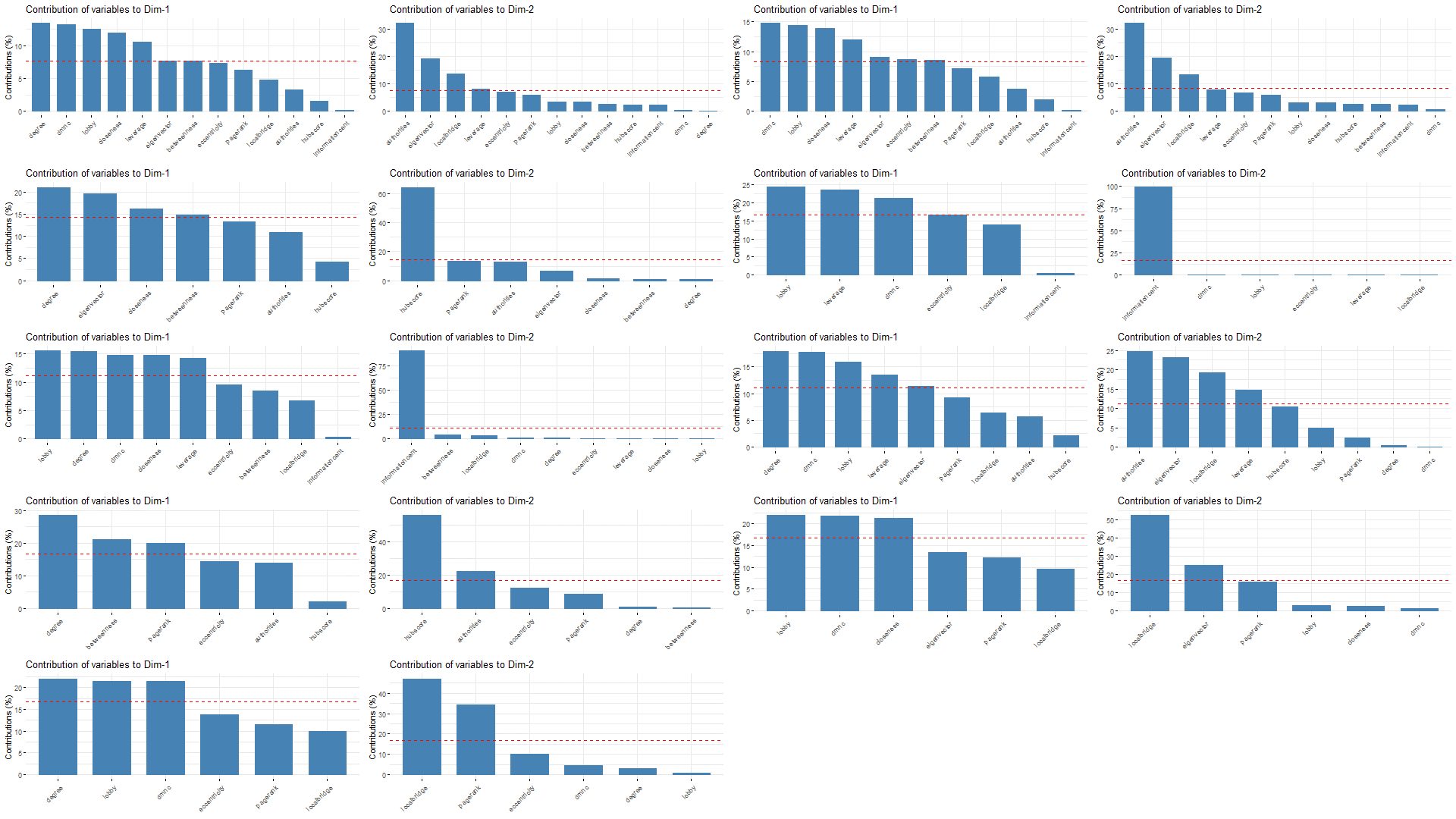
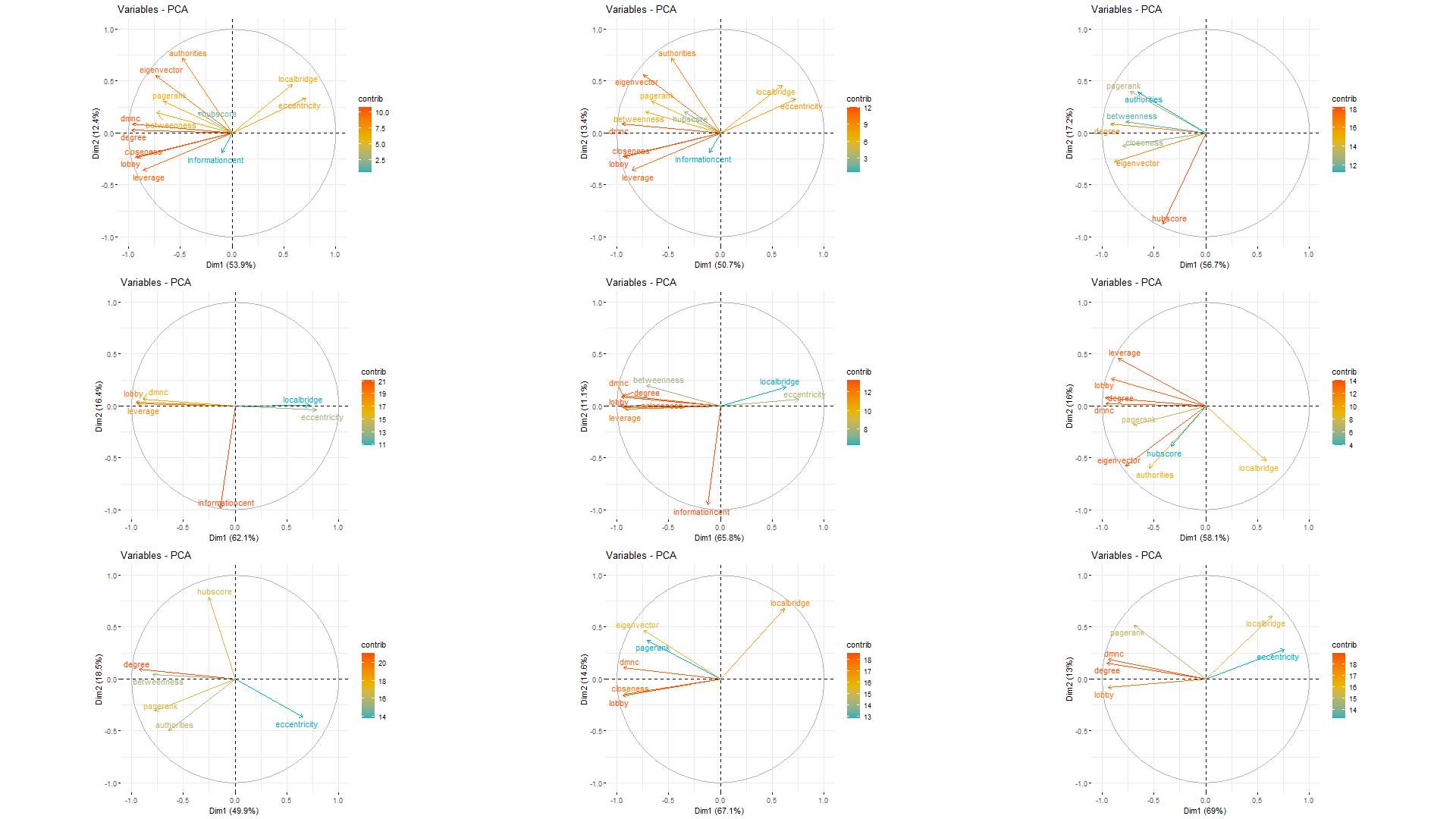


Figure main model

In the main model, the first principal component shows about 53.9% variation, while the first 2 principal components shows about 66.3% variation and finally, taking the first three principal components together capture about 76% variance of the dataset.



From the dimension contribution plots it can be seen that Degree, DMNC, Lobby, Closeness and leverage are the more significant factors contributing to principal component 1. In principal component 2, authorities, eigenvector and local bridging centrality contributes more.

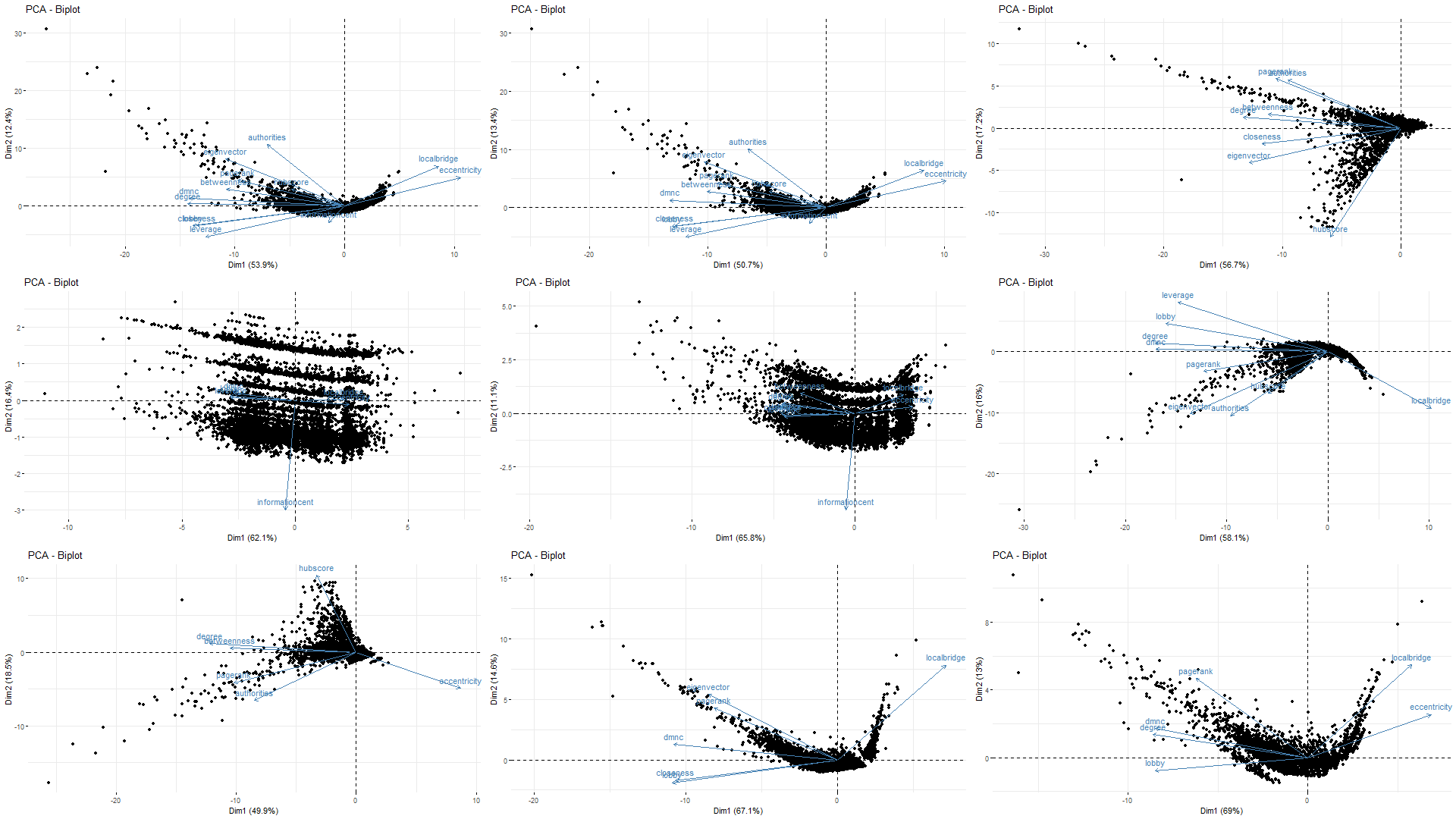


From the cos2 contribution plot it can be seen that the length and color of the variable signifies the importance/contribution of the variables, and hence in the main model degree, DMNC, lobby, closeness, leverage, eigenvector and authorities contribute the most.

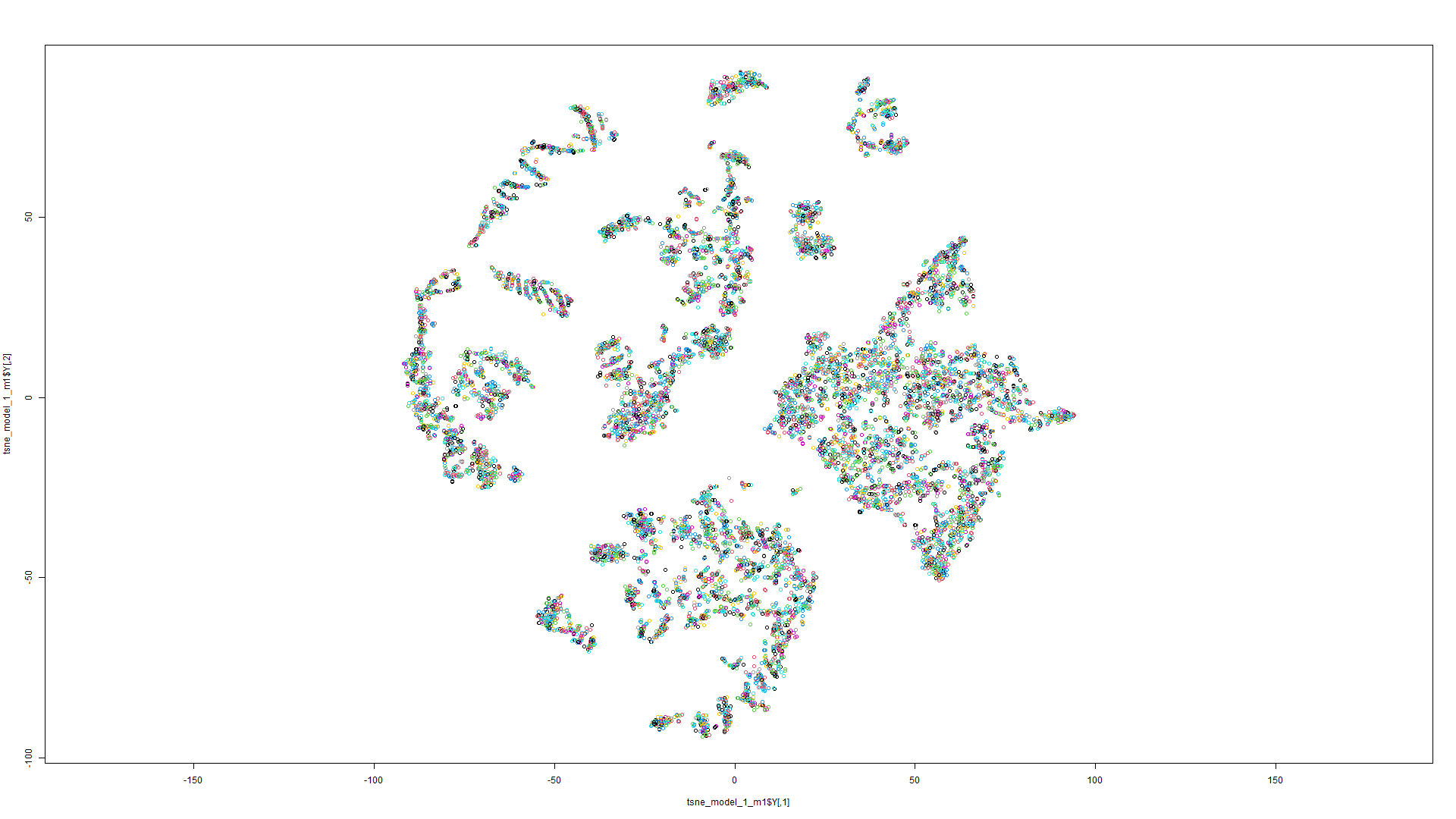
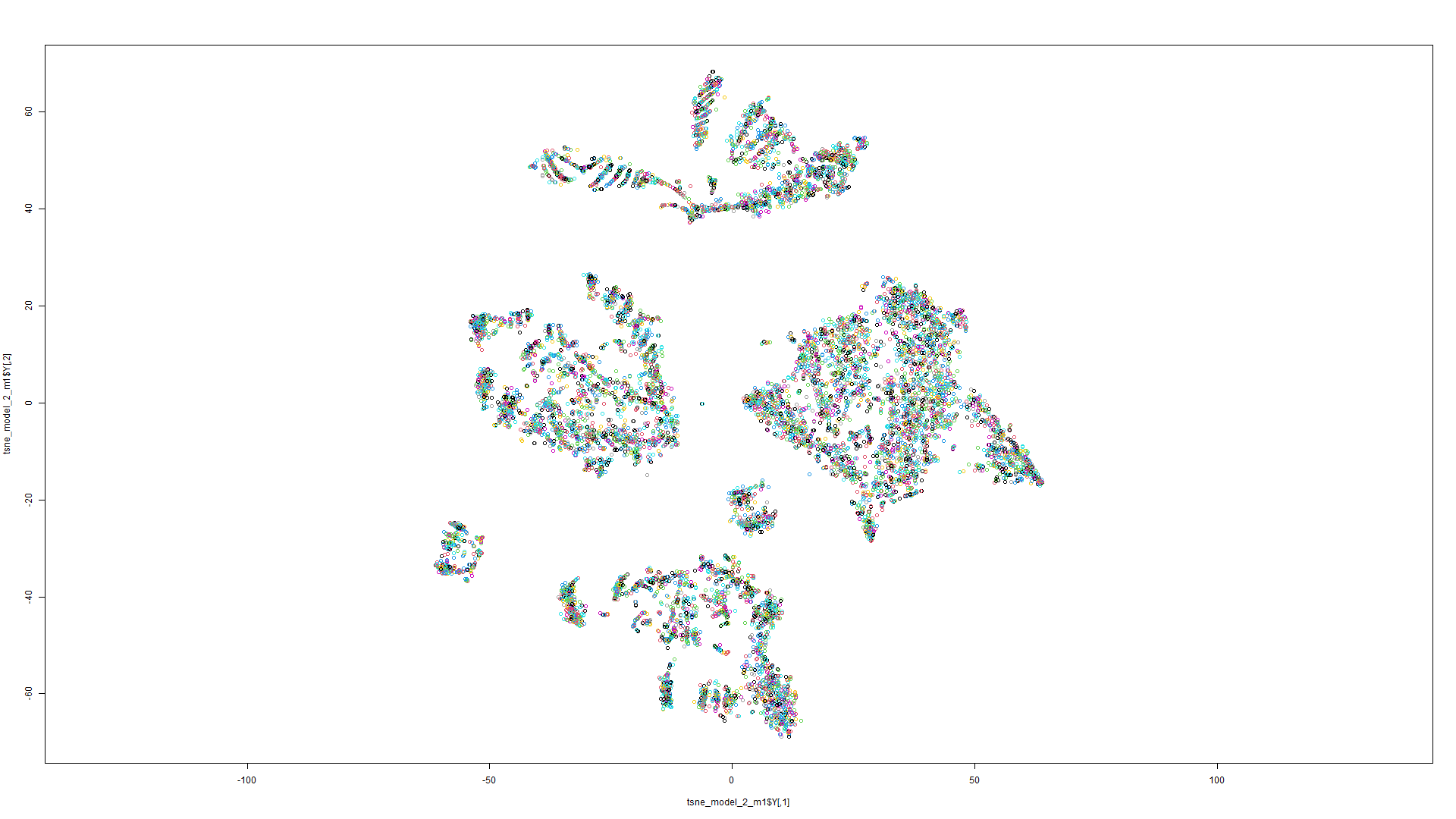
This is consistent with our initial assumptions that degree, leverage and closeness would convey more information for p2p networks. As it has been mentioned before, seeders are more influential in a p2p network. Nodes that seed either a large number of files or a large variety of files are likely to be connected to more nodes, hence they should have higher degree. Such nodes should also have higher leverage values as they have higher degree than their neighbors. As these influential nodes are adjacent to many other nodes, they also have higher closeness values compared to nodes who are less important in the network. Nodes that have direct links with influential nodes are also important as they are likely to share the more in-demand files, such nodes can be expected to seed a subset of the files shared by more important nodes. Therefore nodes that are connected to more influential nodes have higher eigenvector values, causing eigenvector to be an important contributor to the principal components.

However, we also find that DMNC, lobby and authorities are also significant. This makes sense considering nodes adjacent /close to influential nodes can be expected to be in the same component as the influential node, as they will also share parts of the same files with each other. Hence density of the components will be higher if it contains 1 or more influential nodes, hence DMNC should be a factor. The impact of authority centrality can be understood by the following analogy- if a group of influential nodes are directed to a node, then that node should have a high authority value. This can be thought of as these nodes being the first one to upload certain files, for example someone who has special access to those files. These nodes themselves might not have a higher degree due to resource limitations or short time in seeding, however other nodes which have higher degree value will likely download and seed the same files. Hence such nodes can have higher authority values while others do not.

*\*still not clear on how to explain lobby\**



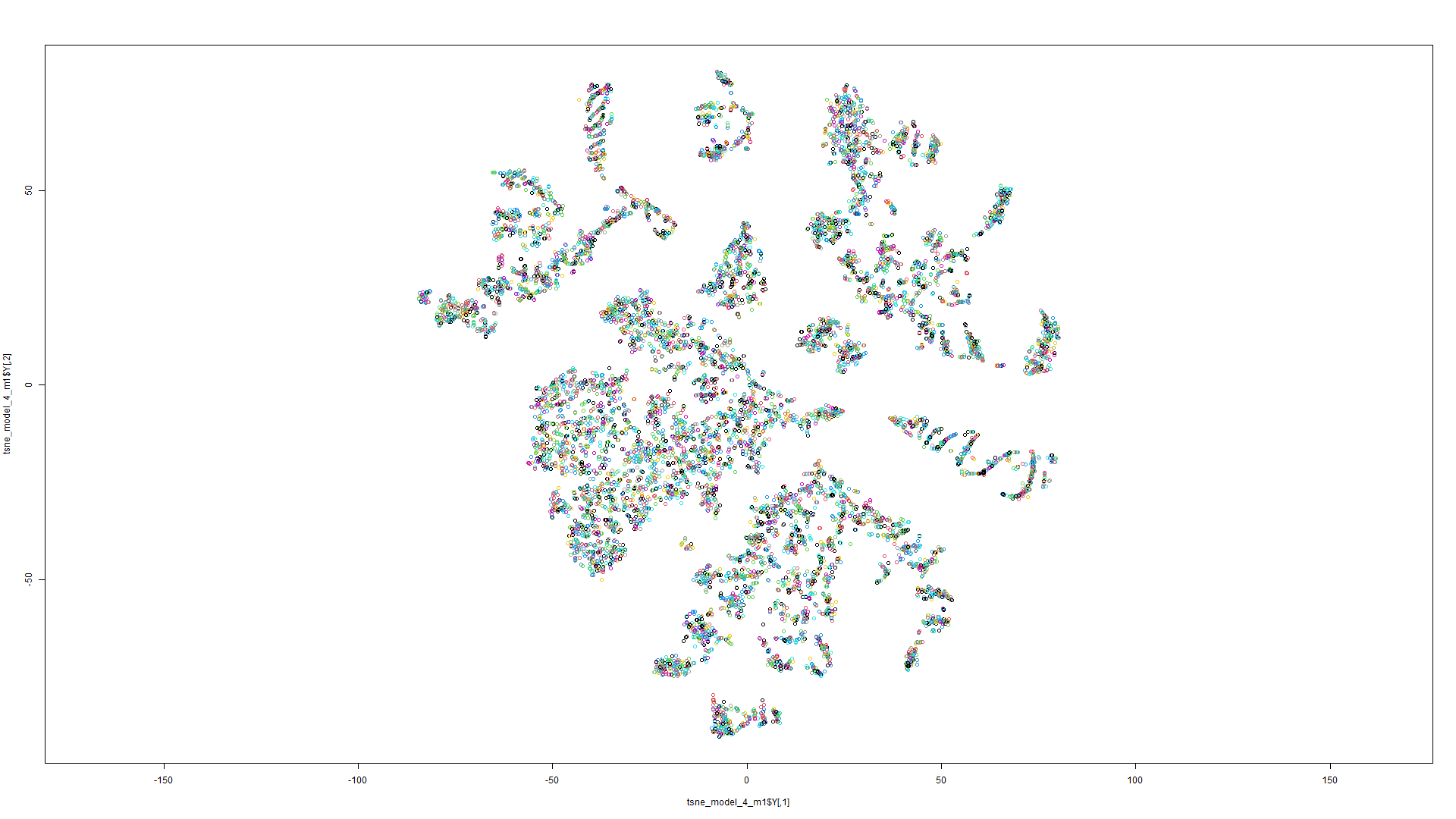
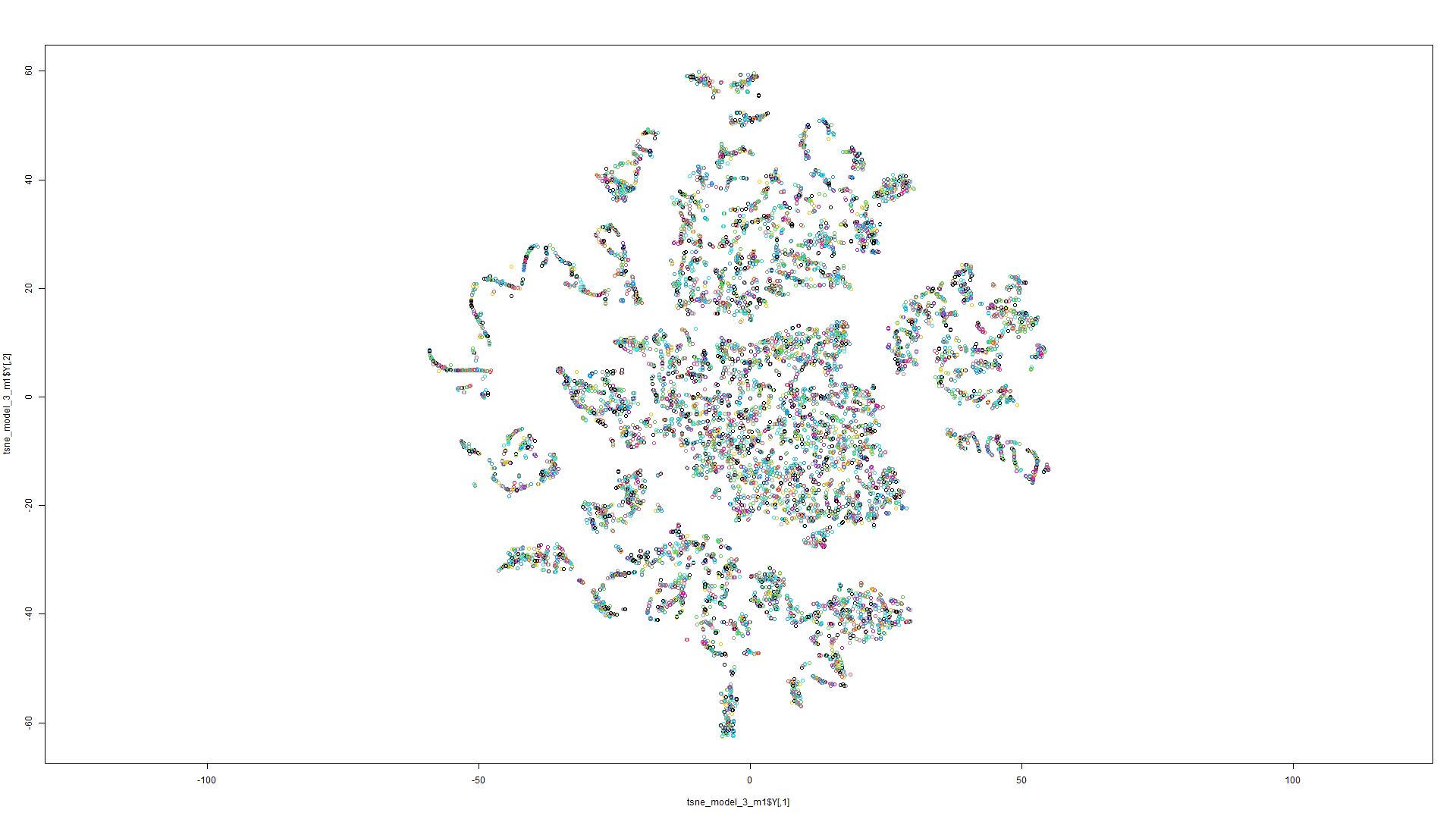
The biplot above is constructed using the PC1 and PC2, which together demonstrated 66.3 variance. From the correlation plot it could be seen that local bridge and eccentricity had negative correlations with the other variables, and this is clearly captured by the principal components as well. Judging by the lengths of the variables, it can be said that most of the centralities except information centrality and Hubscore vary significantly from the average values. However, in contrary to the correlation plot which shows that degree is almost directly positively correlated to DMNC, lobby and leverage, in the biplot it can be seen that while they are strongly associated (due to the small acute angle between them), they are not that strongly associated in the principal components. The concentration of the data points near the origin shows that most data points had similar values when they were projected on these axes, which indicates that there is no strict structure in p2p networks, with aligns with our initial assumption.

Perplexity= 50, Perplexity= 100

Theta= 0.2 Theta=0.1

Iterations=2000 iterations= 1500



Perplexity= 30, Perplexity= 43

Theta= 0.5 Theta=0.1

Iterations=2000 iterations= 150

From the t-SNE models above, it can be seen that some clusters are formed when perplexity is higher. Essentially t-SNE translates the existing clusters from higher dimension to lower dimensions, and the clusters in the 13-dimension dataset are essentially nodes that have very similar combination of centrality values. It is a logical assumption that nodes in a community, or nodes that often interact with each other, have similar centrality values, due to the highly connective nature of P2P networks . Nodes that interact more can be thought of as nodes that have similar interests, such as people who like anime can be one cluster while people who like movies can be another cluster. Keeping a high perplexity value isolates the cluster more strictly, due to its definition.\**I cannot explain the meaning of perplexity very well\** However lowering the perplexity shows the real scenario that the clusters are actually not that isolated from each other. This can be farther proved by the following DBSCAN:

